Pre-training for Speech Translation: CTC Meets Optimal Transport

Phuong-Hang Le Hongyu Gong Changhan Wang Juan Pino Benjamin Lecouteux Didier Schwab

Université Grenoble Alpes Meta

ICML 40 Years International Conference On Machine Learning

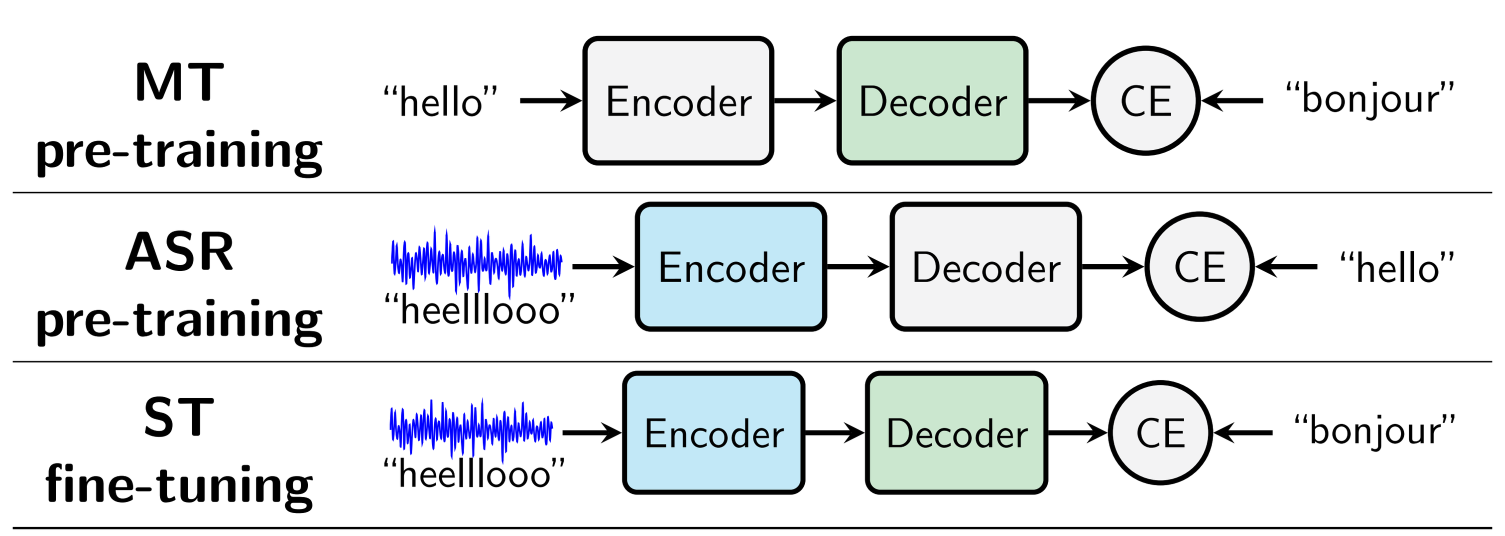
Context and Motivation

- Speech-to-text translation (**ST**): challenging, often requires two auxiliary tasks: automatic speech recognition (**ASR**) and machine translation (**MT**).
- Standard ASR & MT pre-training → **modality gap!**

Contributions

- Showing that connectionist temporal classification (**CTC**) can reduce modality gap.
- New pre-training method: **Siamese pre-training** combining CTC and optimal transport (**OT**).
- Simplicity: our method can reduce modality gap at pre-training stage, requiring no change in ST model.
- Generality: our method can **align sequences of features from different modalities**.

Review of Modality Gap in Pre-training



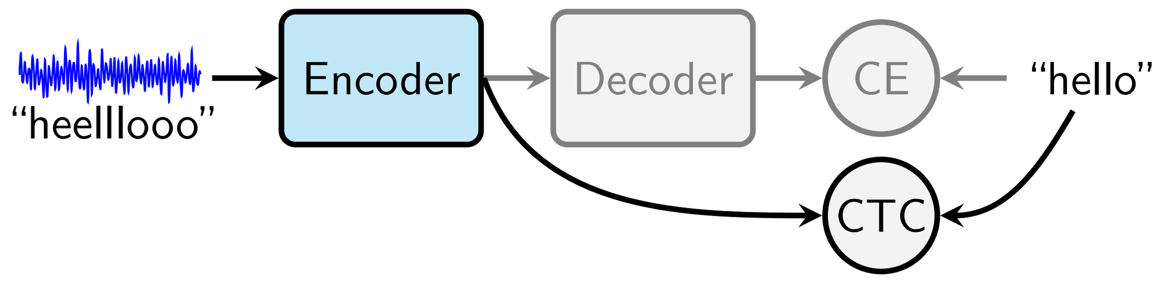
Standard ASR & MT pre-training recipe

ST fine-tuning is initialized with ASR encoder & MT decoder.

CE *stands for cross-entropy.*

✗ Loss of pre-trained *alignment information* due to ASR decoder & MT encoder being discarded during fine-tuning.

Reducing Modality Gap with CTC



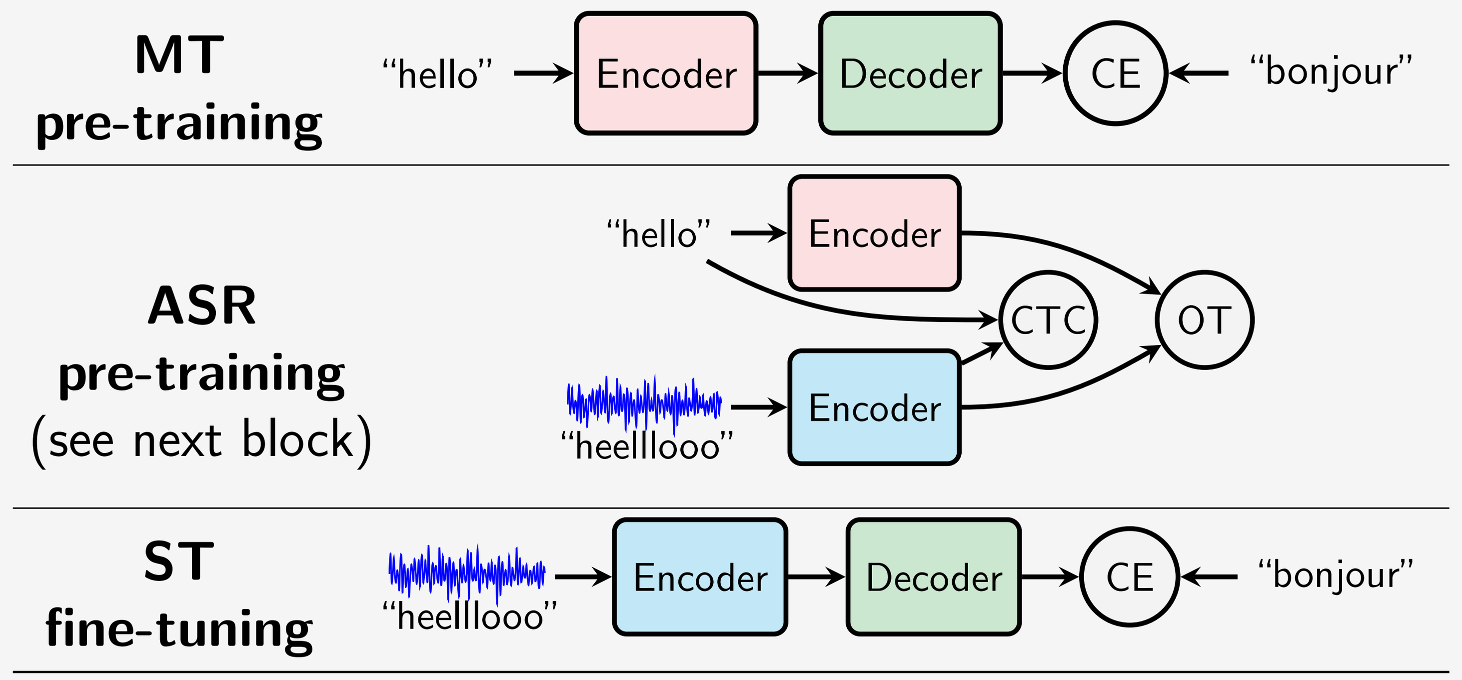
ASR pre-training with CTC. *CE is optional.*

Given audio input $\mathbf{X} \triangleq (\mathbf{x}_1, \dots, \mathbf{x}_S)$ with hidden features $(\mathbf{h}_1, \dots, \mathbf{h}_S) \triangleq \text{ENCODE}(\mathbf{X}; \boldsymbol{\theta})$, CTC [Graves et al., 2006] predicts a text token $\hat{a}_t \in \mathcal{V}$ at each time step t:

$$p(a_t \mid \mathbf{X}) = \text{softmax}(\mathbf{W}\mathbf{h}_t + \mathbf{b})[a_t] \; \forall a_t \in \mathcal{V},$$
$$\hat{a}_t = \operatorname*{argmax}_{a_t \in \mathcal{V}} p(a_t \mid \mathbf{X}).$$

✓ ASR encoder trained with CTC already learns to align speech input to text output without a decoder.

Proposed Siamese Pre-training for ST



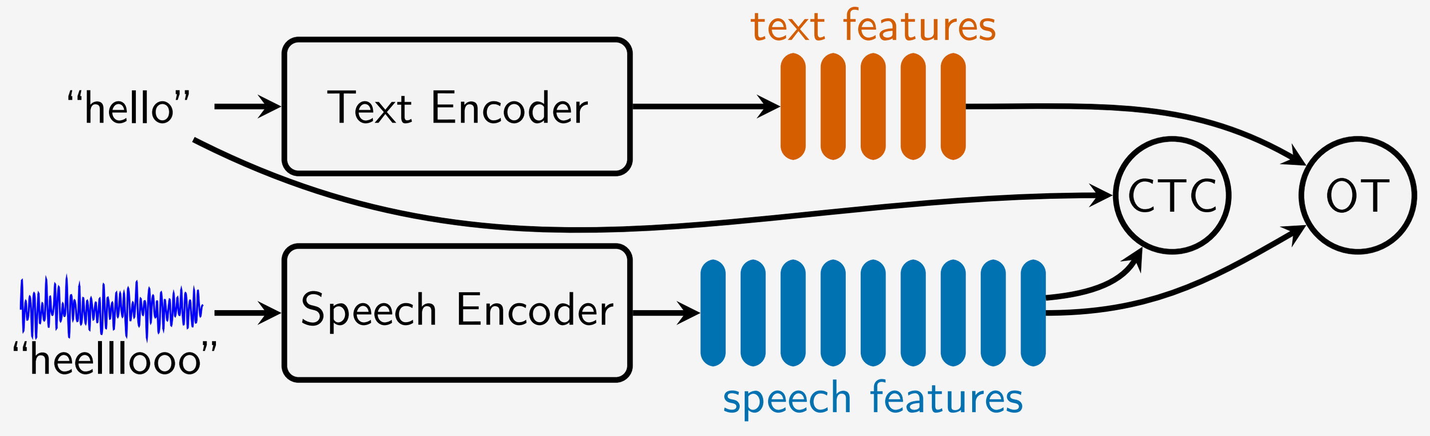
Proposed ASR & MT pre-training recipe

Pre-trained MT encoder is used by OT in ASR step.

✓ All pre-trained components are used.

✓ Optimal transport reduces modality gap by aligning speech and text features.

Optimal Transport for Pre-training



Siamese network for speech-text alignment

OT pulls speech and text features closer in Wasserstein space, while CTC further enhances speech features.

Given speech features $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_m)$, text features $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_n)$ $(\mathbf{u}_i, \mathbf{v}_j \in \mathbb{R}^d)$, and distance function c. The OT (or Wasserstein) loss is defined as:

$$\text{OT}(\mathbf{U}, \mathbf{V}) = \min_{\mathbf{Z} \in \mathbb{R}^{m \times n}} \sum_{i=1}^{m} \sum_{j=1}^{n} Z_{ij} c(\mathbf{u}_i, \mathbf{v}_j),$$
$$\text{s.t.} \quad \mathbf{Z} \geq \mathbf{0}, \; \sum_{j=1}^{n} Z_{ij} = \frac{1}{m}, \; \sum_{i=1}^{m} Z_{ij} = \frac{1}{n}.$$

Interpretation: OT finds the transportation plan $\mathbf{Z}$ with minimum cost between two distributions.

- $\mathbf{U}, \mathbf{V}$: mass locations of two uniform distributions.
- $c(\mathbf{u}_i, \mathbf{v}_j)$: unit cost of transporting from $\mathbf{u}_i$ to $\mathbf{v}_j$.
- Z_{ij}: quantity of mass transported from $\mathbf{u}_i$ to $\mathbf{v}_j$.

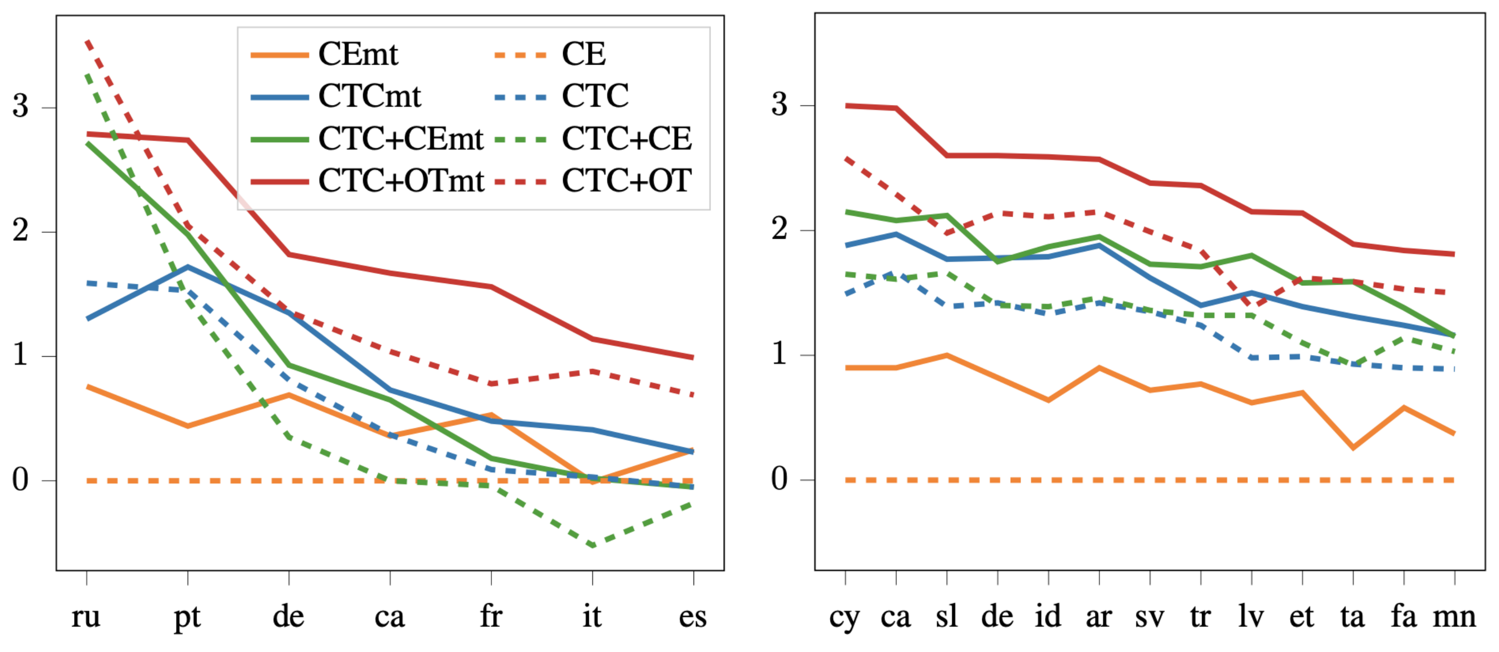
Positional encoding for OT

Motivation: OT loss ignores sequence orders, while our encoder inputs are *monotonically* aligned.

Idea: integrating normalized positions $s_i = \frac{i-1}{m-1}$ and $t_j = \frac{j-1}{n-1}$ into cost function:

$$c(\mathbf{u}_i, \mathbf{v}_j) = \left(\|\mathbf{u}_i - \mathbf{v}_j\|_p^p + \gamma^p |s_i - t_j|^p \right)^{1/p}.$$

Experimental results



(a) CoVoST-2 X*→En. (b) CoVoST-2 En→X.

Results on CoVoST-2

BLEU relative to CE. "mt" means MT pre-training was performed.

Method		Multi	BLEU de	es	fr	it	nl	pt	ro	ru	avg
FAIRSEQ S2T [Wang et al., 2020]		✓	24.5	28.2	34.9	24.6	28.6	31.1	23.8	16.0	26.5
ESPnet-ST [Inaguma et al., 2020]		✓	22.9	28.0	32.7	23.8	27.4	28.0	21.9	15.8	25.1
Dual-decoder [Le et al., 2020]		✓	23.6	28.1	33.5	24.2	27.6	30.0	22.9	15.2	25.6
Adapters [Le et al., 2021]		✓	24.7	28.7	35.0	25.0	28.8	31.1	23.8	16.4	26.6
BiKD [Inaguma et al., 2021]		-	25.3	-	35.3	-	-	-	-	-	-
JointSpeechText [Tang et al., 2021]		-	26.8	31.0	37.4	-	-	-	-	-	-
TaskAware [Indurthi et al., 2021]		-	**28.9**	-	-	-	-	-	-	-	-
ConST [Ye et al., 2022]		-	28.3	32.0	38.3	27.2	**31.7**	33.1	25.6	**18.9**	29.4
STPT [Tang et al., 2022]		-	-	**33.1**	**39.7**	-	-	-	-	-	-
CE pre-training	LARGE	✓	26.9	30.8	37.7	26.7	30.8	33.3	26.2	17.9	28.8
CTC pre-training	LARGE	✓	27.6	31.4	38.2	27.2	31.1	33.6	26.4	18.4	29.2
CTC+CE pre-training	LARGE	✓	27.2	31.2	38.0	27.0	31.5	33.7	26.2	18.3	29.1
Siamese-PT (this work)	LARGE	✓	27.9	31.8	39.2	**27.7**	**31.7**	**34.2**	**27.0**	18.5	**29.8**

Results on MuST-C

Main takeaways

- Encoder trained with CTC is stronger than the one trained with encoder-decoder-CE.
- Our Siamese pre-training helps reduce modality gap without any changes in the ST model.
- Optimal transport is very effective for learning to align sequences of features from different modalities.